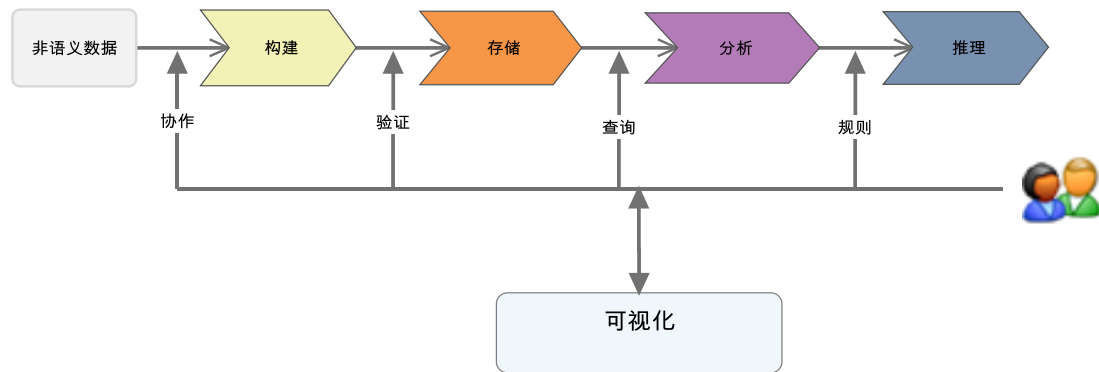


# 语义网可视化(Visualization for Semantic Web)

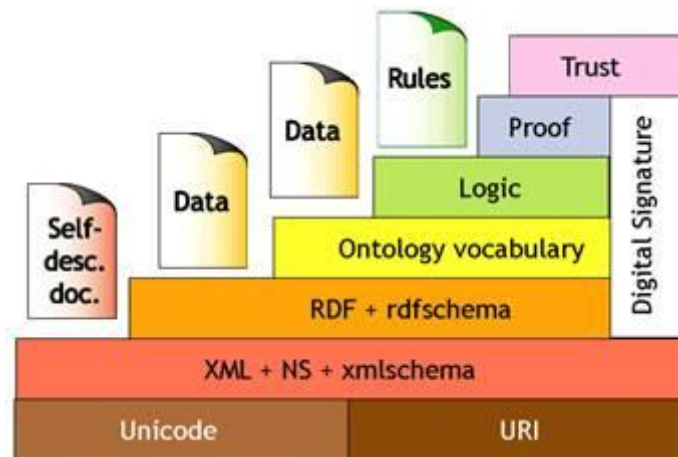


## 1. 语义网

语义网 (Semantic Web) 是由万维网联盟的蒂姆·伯纳斯-李(Tim Berners-Lee)在 1998 年提出的一个概念，它的核心是：通过给万维网上的文档 (如 HTML)添加能够被计算机所理解的语义(Meta data)，从而使整个互联网成为一个通用的信息交换媒介（维基百科）。其最基本的元素就是语义链接(linked node)。

### 1.1. 语义网的结构

蒂姆·伯纳斯-李创造性的将 XML, RDF, 本体论技术的应用结合起来，设计出了语义网的技术层次架构图，如图 1-1，是应用最多的关于语义网结构的层次模型，或称协议栈（Protocol Stakes）。



**Unicode 和 URI 层：** Unicode 用来定义国际化、通用化的字符集，URI 是统一资源标识符的缩写。语义网中，所有事物都称为资源，而每一个资源都用唯一的 URI 标识。

**XML+NS+xmlschema 层：** 语义网的基础描述语言层。XML 是语义网层次模型的基础，命名空间（Name Space）为 XML 文档中的结构化标记的定义和使用提供上下文机制，用以指明涵义，避免命名冲突。XML Schema 为 XML 文档提供了语法结构上的约束，保证 XML 文档的完整性与有效性。

**RDF/RDF Schema 层：** RDF 资源描述框架（Resource Description Framework）是一种用于描述网上资源的语言。RDF Schema 则为 RDF 更丰富的结构表达提供了一套类型定义系统。

**本体（Ontology）层：** 本体技术是语义网的核心。本体的功能就是提供网上互操作体之间关于信息的共同理解，也就是“语义”。

**逻辑（Logic）层：** 该层用以提供公理和推理规则，为智能推理提供基础。

**证明（Proof）层：** 证明层用于提供认证机制，执行逻辑层产生的规则。

**信任（Trust）层：** 主要负责提供信任机制，保证资源的交互安全可靠。

## 1.2. 语义网基础

语义网中最重要的两个基础是 **RDF** 和 **本体**。在语义网的交互中，本体担当着语义互操作的重要角色。实践中，用 **RDF** 定义了网上信息资源，再用本体定义互操作的语义空间，就构成了一个基本的语义网应用环境。加入规则(rule)子层，可以提高本体描述能力，增强信息的语义表达能力。规则子层还可以定义与具体应用相关的知识描述，提供个性化的私有描述。

**RDF** 基于这样的思想：用 **Web** 标识符（**URIs**）来标识事物，用简单的属性（**property**）及属性值来描述资源。这使得 **RDF** 可以将一个或多个关于资源的简单陈述表示为一个由结点和边组成的图（**graph**），其中的结点和边代表资源、属性或属性值。**RDF** 采用“主体—谓词—客体”的三元组（**Triples**）方式描述资源。

哲学上，本体能大致对应的概念是范畴，来源于柏拉图和亚里士多德的观点：本体是研究事物之所以存在的科学。现代人工智能对本体的定义各不相同。总结这些定义的共同点，**Fensel** 教授将其整理为本体的四大特征：

**概念化（conceptualization）：** 将客观世界中的现象抽象成概念模型。

**形式化（formal）：** 用精确的数学语言而非自然语言描述这些概念，本体是计算机可读的。

**明确的（explicit）：** 本体中的概念及其关系都是明确的定义的。

**共享（share）：** 本体反映的是相关领域中公认的概念集，目的是为了知识的共享。

本体描述语言可分为三类。

**基于逻辑演算的本体描述语言：** 这类本体描述语言都是基于某种逻辑系统设计的，且语法具有明显的逻辑特征。和纯粹的逻辑系统相比，它们的语法又具有少量的自然语言的成分，但可读性较差。

**基于图的本体描述语言：** 这类语言除了底层的支持逻辑系统以外，还有专门的图表示层，即有一个从图结构到底层逻辑结构的双射集合。同基于逻辑演算的语言相比，基于图的本体描述语言最大的优势就是对人的可读性好，但其对机器的可读性较差，此外对于复杂关系的表达能力比较弱。

**基于分布式环境的本体描述语言：** 分布式环境下的本体描述语言的最大特点是增加了基于

标签语言的文本表示层。基于标签语言的本体描述是为了适应分布式环境下的知识描述特点，实现网络环境下的信息互享，数据互操作。基于标签的文本描述特别有利于解析本体文档，因而对机器的可读性较好。因为标签语言近似自然语言，其对人的可读性也要好于基于逻辑演算的描述语言。目前 W3C 官方本体描述语言是 OWL，兼容 RDF。

## 2. 语义网的构建

本体作为语义网的知识表示模型很好地解决了语义层次上 web 信息共享和交换问题。构建语义网需要有丰富且实时更新的本体。因此如何创建和管理本体是实现语义网的基础。

### 2.1. 构建工具

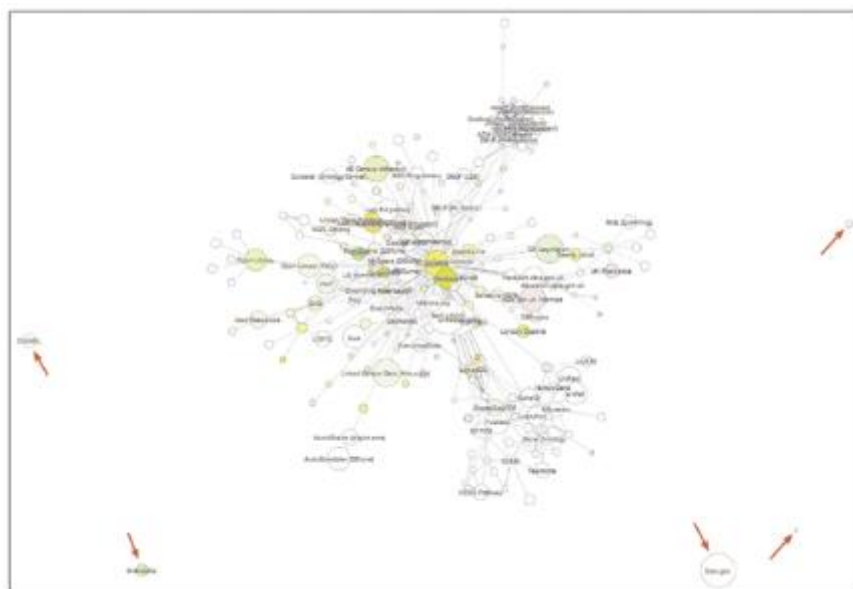
目前本体的构建工具包括：Ontolingua, Ontosaurus, WebOnto, WebODE, OntoEdit, OILEd, Protégé 等。下表是关于这些工具的对比。

工具名称	Ontolingua	Ontosaurus	WebOnto	WebODE	OntoEdit	OILEd	Protégé
OWL 语言	x	x	x	√	x	√	√
中文支持	x	x	x	x	x	x	√
可视化	x	x	√	√	x	x	√
网络技术	√	√	√	√	√	x	x
合作开发	√	√	√	√	√	x	x
本体合并	√	x	x	√	√	x	√
模糊本体	x	x	x	x	x	x	x

从表中可以看出，在可视化、中文支持、本体合并和模糊本体支持上仍存在不足，特别是模糊本体。构建本体过程中，知识和信息有时候是不确定的，如何对现有的本体进行拓展构建模糊或近似本体，这对于解决领域中的不确定性推理问题非常重要。此外，随着处理的知识规模不断增长，构建本体需要不同的领域专家共同协作。这些都为本体可视化带来了一定的机遇。

由于本体构建是一项繁琐的任务，自动和半自动的构建方法成为研究热点。Shamsfard 等提出手动构建核心本体，在此基础上利用基于文本的学习技术自动扩充本体。Khan 等从字典中抽取感兴趣的概念和关系，构建需要的本体。Celjaska 等提出一种对非结构化文档构建本体的方法，基于监督学习，从已标记文档中提取规则，将所得规则应用于新文档集合构建本体。

自动或半自动工具的使用减少了人工参与的工作量，同时带来的问题是构建的结果可能并不令人满意，其中存在大量的错误。下图边界地区出现了一些离群结点，说明这些本体的构建存在一些错误。在这种情况下，发现并调整构建结果作用非常明显。



## 2.2. 语义识别

**语义标注：**采用 RDF，Microformats 等语法标注原数据——最正统的方法。这种方法可看成是给词打上对应的语义标签，其目的主要是为了解决数据的互通性，比如交友社区中的朋友关系如果使用语义标注，就可帮助全网（跨不同网站的）范围内的朋友搜索（具体可参见 <http://www.ibm.com/developerworks/cn/xml/x-watch/part3/index.html> ）。让语义网如此有意义和有用的关键设计特征之一就是跨 Web 数据集间互连数据的确立。就这点而言，目前最成功的项目我认为是 Open Linked Data，它已经整理了不少领域的的数据信息可供识别。

**建立语境：**这种方法是通过有向图的方式描述我们身边的客观世界。网络上的节点表示物理实体、概念或状态，连接节点的“边”用于表示实体的关系。通过边的链接实体间的相关性得以表示，因此相关事实可以从直接相连的节点推导出来——而且目前语境的建立多采用面向对象中的属性继承方式，因此可以对继承的属性进行演绎如三段式推理，另外甚至可以建立状态和动作的描述（后面提到的 freebase 使用类似方法）。有了这种网络描述的语境（和专家系统有点像），那么就能够判断文档内容是何含义了——最简单的方法是将文档的词汇抽取出来，在各种语境网格下演算其匹配程度，最后收敛到最匹配的语境就代表了该文章的含义。

**聚合方法：**相比建立语境技术，聚合所采用的是基于统计的词群聚类。一般使用监督模型训练好不同语义的词群，然后分析文章的词汇主要落在那个词群中，就说明它的语义属于那个词群对应的语义。最后识别语义方式和语境下识别有相似之处，不同之处是：前者在语境图中进行匹配，后者则是在匹配词群中匹配。

然而，正如前文所述，标注和语境都需要大量的人工干预，而聚合方法虽是自动化的，但分

类比较粗，而且计算量较大。自然语言处理技术在一定程度上能够解决语义网构建问题。

### 3. 语义网的存储管理

很多本体之间存在明显的继承关系，树可以用来描述本体。然而由于不同本体类别之间存在某种联系，图模型描述本体更加自然。

#### 3.1. 图数据的存储

图模型的常用存储结构包括邻接矩阵、邻接表、十字链表和邻接多重表。从大规模图处理的应用需求和维护的复杂度上考虑，邻接矩阵和邻接表更具有优势。

大规模的图数据存储需要依赖云计算环境的分布式存储系统。云计算环境的存储系统分为两种：一种是以 GFS、HDFS 为代表的分布式文件系统，对于邻接矩阵、邻接表等结构，可以直接存放；另一种是以 BigTable、HBase 为代表的 NoSQL (NotOnlySQL) 分布式数据库。NoSQL 数据库采用的数据模型主要有文档储 (Document Store) 模型、列族存储

(Column-Family Store) 模型、Key-Value 存储模型、图存储模型几大类。与其他三种模型相比，key-value 更适合于存储大规模的图数据。Key-Value 存储模型的存储模式简单，支持海量数据存储和高并发查询操作，非常适合通过主键进行查询或遍历，但对复杂的条件查询支持度不佳。

语义网本体描述为三元组，Key-Key-Value 模型存储的信息比传统的 Key-Value 模型更加丰富，可以据此进行数据迁移和合并，以提高空间局部性，使得在查询处理时能减少远程读取数据的次数，进而提高数据的读取效率。

国内华中科技大学服务计算技术与系统实验室开展的海量数据管理项目：大规模 RDF 数据管理系统 TripleBit。

#### 3.2. 图数据的索引结构

B+树，R 树等

#### 3.3. 图数据的分割

对于一个大规模图的处理，必须进行分布式并行处理。由于图数据本身固有的连通性和图计算表现出强耦合性的特点，为了实现高效的并行处理，尽可能降低分布式处理的各子图之间的耦合度是非常重要的。有效的图分割就是实现解耦的重要手段。

将一个大图分割为若干子图，有两个主要原则：一是提高子图内部的连通性，降低子图之间的连通性，这种特点尤其适合云计算的分布式并行处理机制；二是考虑子图规模的均衡性，尽量保证各子图的数据规模均衡，不要出现较大的偏斜，从数据规模方面防止各并行任务的执行时间相差过大，降低任务同步控制过程中“水桶效应”的影响。

如果只考虑数据负载均衡这一单项指标，最简单的图分割技术，就是 Hash 方式，即在设定了分片数目之后，对图顶点 ID 进行 Hash，将数据划分成给定数目的分片。这种分割方法效率很高，时间复杂度为  $O(n)$ ，可以在图数据的载入过程中或图处理之前完成分片操作。

如果只考虑子图内敛性这一单项指标，即增大子图内部的关联性，降低子图之间的关联性，可采用聚类技术，效果十分明显。Apache 开源项目 Mahout 在云计算环境下实现了分布式聚类方法。Yahoo 研究院开发出“Local Partition”算法，该算法的运行时间与最终输出结果的聚簇大小成正比，而与图的原始输入数据规模无关，从而可以对更大规模的图进行分割处理。

同时考虑子图数据规模均衡和子图内敛性等多项指标，也有很多研究者进行了尝试，但是应用都受到不同程度的限制。

#### 4. 语义网分析

语义网的分析主要体现在两个方面，一是文档层次的分析，二是对文档内容语义的分析。文档级的分析例如语义网文档来源网站的主要顶层域名后缀分布，以及语义网文档的年龄分布；swoogle 通过计算语义网文档 RDF 图的三元组个数，来分析语义网文档的大小分布。近些年来，由于微博等社交网络的发展，语义网的分析大量应用于社会网络。

Boanerges 等利用语义网技术探测社会网络内的兴趣冲突，以及社会网络内数据的获取和歧义消除。通过语义社会网找到人与人之间的联系，测量合作的力度。Thomas K. Yan 等研究如何将社会网络的信息用于加速建立可靠的元数据，模拟不同信任级别合作情况。John C. Paolillo 等提出语义网可视化模型，运用社会网络分析算法对 LiveJournal FOAF 上的数据进行了分析。

SIOC 和 FOAF 是语义网在社会网络领域的两种典型应用模型。SIOC(Semantically Interlinked Online Communities)提供给人们互相之间讨论的方法如博客、论坛和邮件列表等。它由一个语义互联在线社区本体、一个表达互联上显性和隐性信息开放标准组成，元数据来源于大量的博客系统和内容管理系统，目标是为用户存储和浏览、查找信息。FOAF(Friend of a Friend)是基于 OWL 的词汇来描述个人信息和个人社会网络的本体模型。FOAF 通过使用 OWL 词汇来形容个人特征，表达能力上已超过了图(raw graph)描述语言。更重要的是，通过使用 FOAF 可以凭借 RDF/OWL 的扩展性来提高特定身份本体查找兴趣相投的人。

分析语义网，其基础是语义查询。由于 RDF 图中存在语义信息，所以传统的图搜索并不完全适用于语义网。SPARQL 是针对 RDF 数据查找的查询语言，类似于 SQL 能够根据用户指定的条件匹配精确信息。对于普通用户使用 SPARQL 是比较困难的，语义网浏览是一种较好的查询方式。语义网浏览提供某种简便交互方式让用户指定兴趣区域 (ROI)，而不需要

输入复杂的查询语句就能获得查询结果。

5. 语义网推理

(待整理)

6. 可视化在语义网中的应用

语义网可视化方法必须能表示本体各单元如类(或实体类型)、关系、实例和属性(或 slots)。其中关系主要有继承关系、包含关系和角色关系(role relations)。继承又分为单继承(分类关系, is-a relations) 和多继承。包含关系是指类与属性之间的关系。角色关系则是属性之间的关联关系。可视化中, 最难表示的多重继承和角色关系, 连线较多、标签复杂。

树和图是语义网可视化最基本的表示方法。树在文件系统和分类文档的管理中使用非常广泛, 在某种程度上, 本体具有类似的特性, 如下表所示。

File system objects	Categorized documents	Ontology
Folder	Category	Entity (class or instance)
Folder/subfolder relationship	Category/subcategory relationship	isa-relationship
Tree view	Categorization	Taxonomy
File	Document	Instance
File properties	Document properties	Slots

6.1. 语义网可视化方法

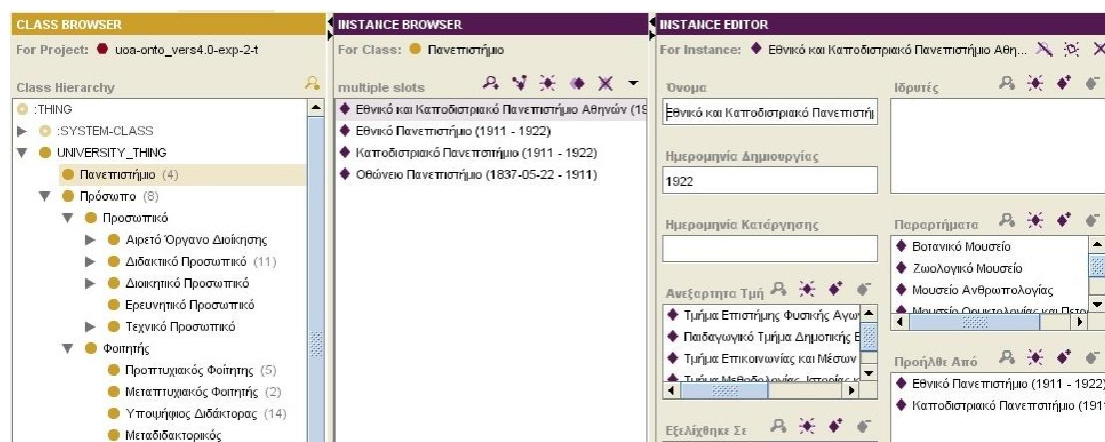
总体来说, 语义网可视化方法可分为以下几类:

- 1) 缩进列表(Indented List)
- 2) 结点链接图和树(Node-link and Tree)
- 3) 缩放表示(Zoomable)
- 4) 空间填充(Space-filling)
- 5) 焦点+上下文和变形(Focus+Context or Distortion)
- 6) 三维信息地形(3D Information Landscapes)

6.1.1. 缩进列表

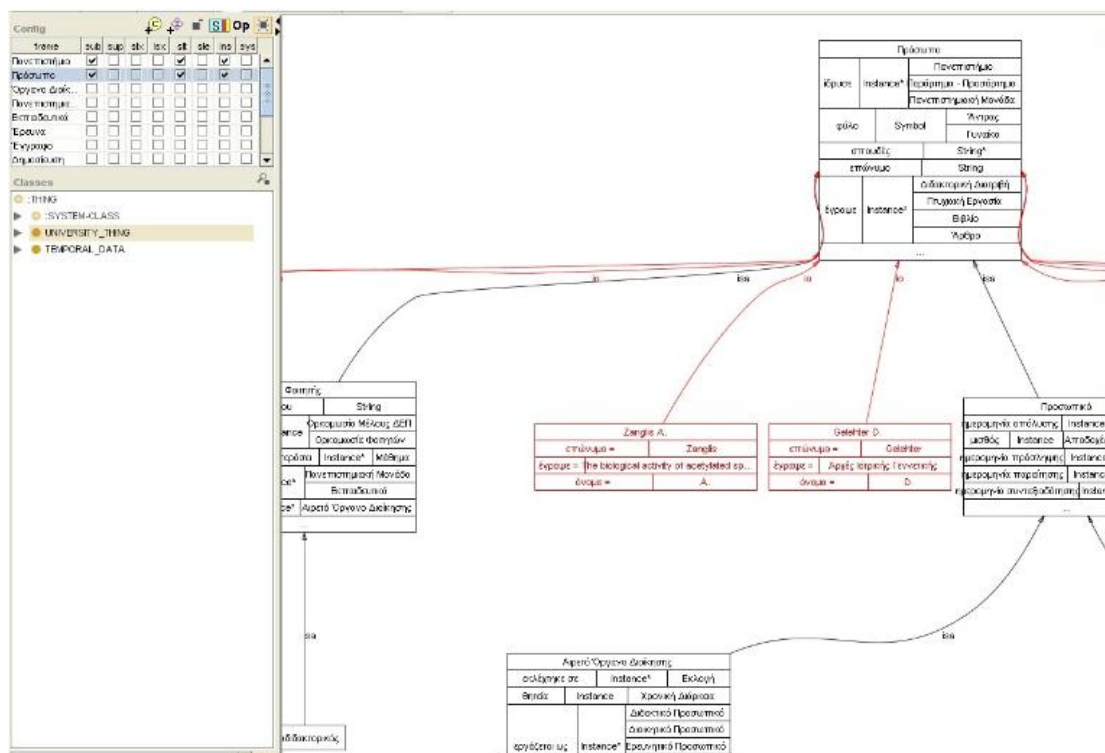
Protégé, OntoEdit, Kaon 和 Ontorama 这些代表性的本体可视化工具都提供树形结构的缩进列表来浏览本体。缩进列表提供关键字查找, 能很好地描述类之间的继承关系和类对实例的

包含关系，但是对于角色关系无法直接表示。



### 6.1.2. 结点链接图

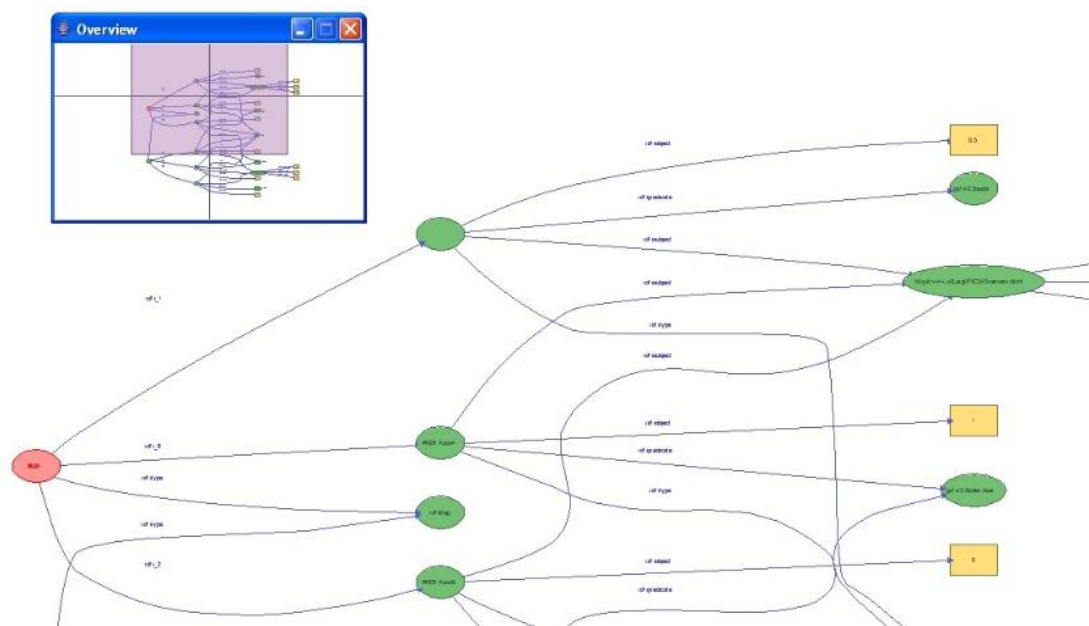
结点链接图将本体表示为相互连接的结点，以自顶向下或从左到右的布局显示分类。用户通常可以展开或收起结点或子树，来调整显示的信息避免视觉混乱。下图是 OntoViz(Protégé 的插件)可视化结果，图中本体的类、属性、继承关系和角色联系都能表示。实体用不同颜色区分，用户还可以通过左上侧的面板选择感兴趣的特征，右键可以缩放。



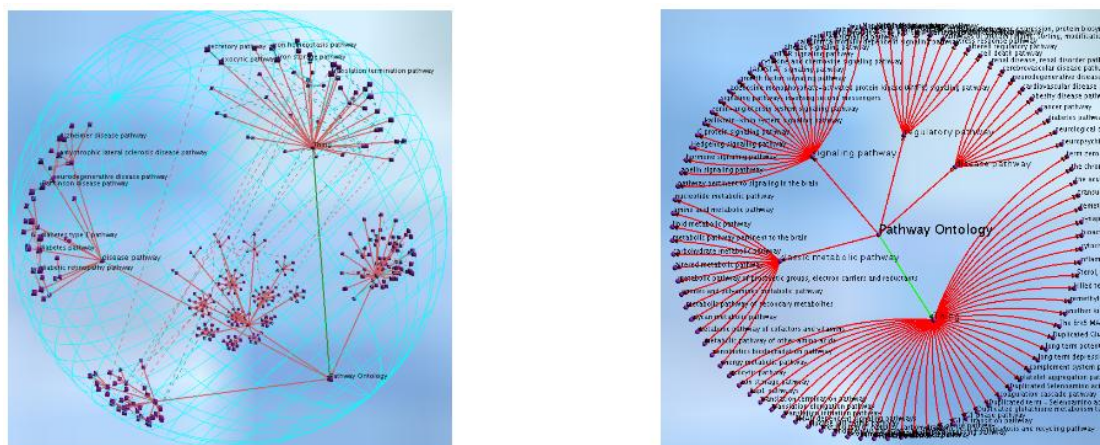
另外，IsaViz，SpaceTree，TreePlus，OntoTrack，GoSurfer，GOBar 也都是用结点链接图表示本体的可视化工具。

GoSurfer 是一个数据挖掘工具，能够对输入的特定基因本体序列(Gene Ontology Consortium, GO)进行可视化。使用自顶向下的树可视化工具将基因与对应的基因本体对比，如比较本体路径，如下图所示。





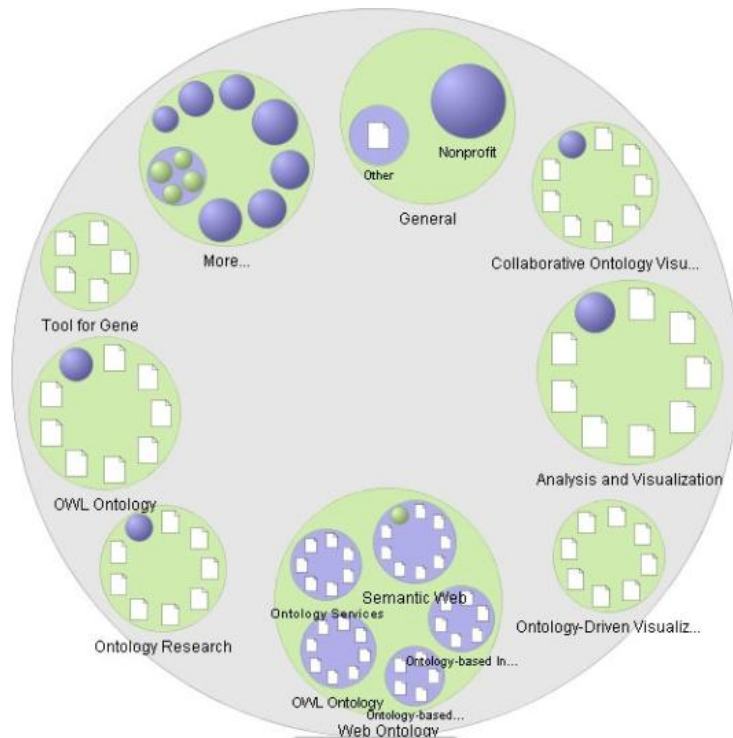
Traversal View 被认为是本体中自包含部分。通过定义 traversal path，可以将链接图收放到用户感兴趣的级别，可以减少视觉混乱。下图为多视图可视化结果，用户可以选择感兴趣的属性列表，从中心概念到属性的传递路径便能显示出来。



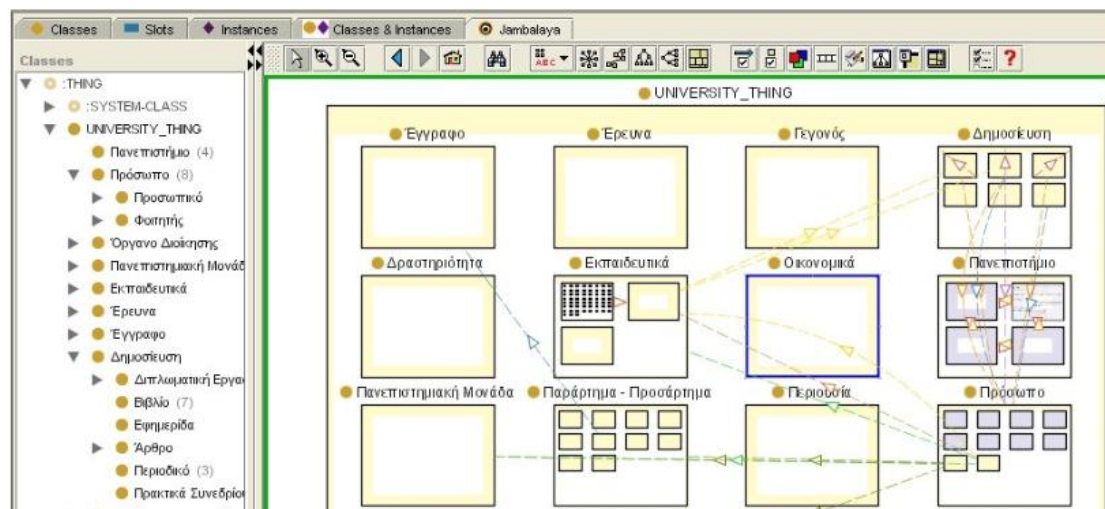
### 6.1.3. 缩放表示

这种可视化方法让子级结点嵌入在父级中，父级结点尺寸要大于子级节点。用户可以 zoom-in 到子结点来放大信息。

Grokker 是一个知识图表示系统，提供了一般化的信息搜索和文件查找的图形化表示。文档聚类后表示成嵌套的 Venn 图。用户可以点击环进入到某一层，当文档被选定后，其内容将在另一个窗口中显示，如下所示。



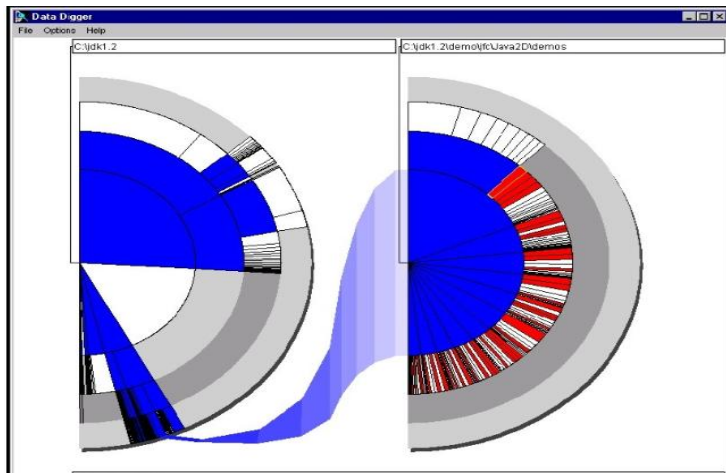
Jambalaya 使用 SHriMP(Simple Hierarchical Multi-Perspective)可视化技术，将本体表示成嵌套可变视图，如下图所示。



#### 6.1.4. 空间填充

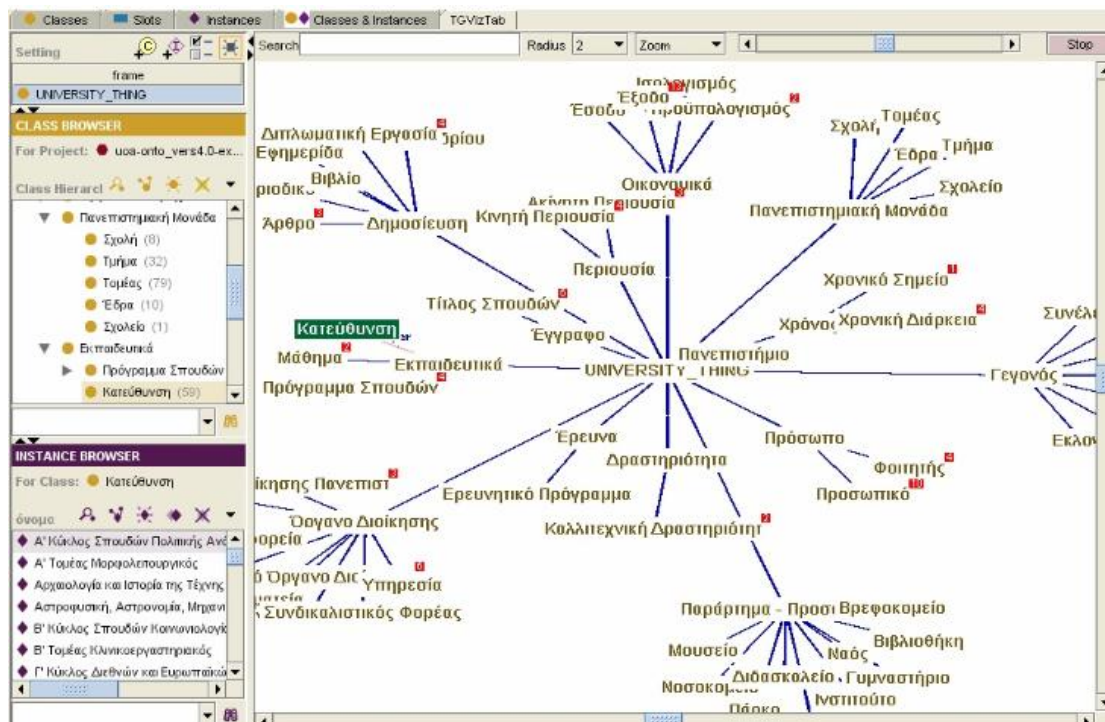
空间填充技术基于使用整个屏幕空间的思想，将空间分割成可以容纳节点的子空间。每次分割的大小取决于节点的属性如节点大小、包含的子节点数等。

树图（TreeMap）是典型的空间填充技术，用嵌套的矩形区域表示节点间的层次关系。信息切片(Information Slices) 用多个半圆盘可视化紧凑的多层次结构。每个盘片表示多个层次，一般是 5-10 级，用户也可以调节。子节点使用可用的分割空间，详细的节点信息还可以通过另外的窗口扩展显示，如下图。



### 6.1.5. 上下文+焦点

上下文+焦点的方式通过对当前视图的变形将上下文和焦点组合到一起。焦点处的节点通常是中心，其他节点尺寸逐级减少环绕在周围，直到不可见为止。用户可以指定不同节点为焦点对其放大。下图显示的是 TGVizTab 可视化结果。其他的还有 OntoRama, MoireGraphs, Bifocal Tree, OZONE 等。



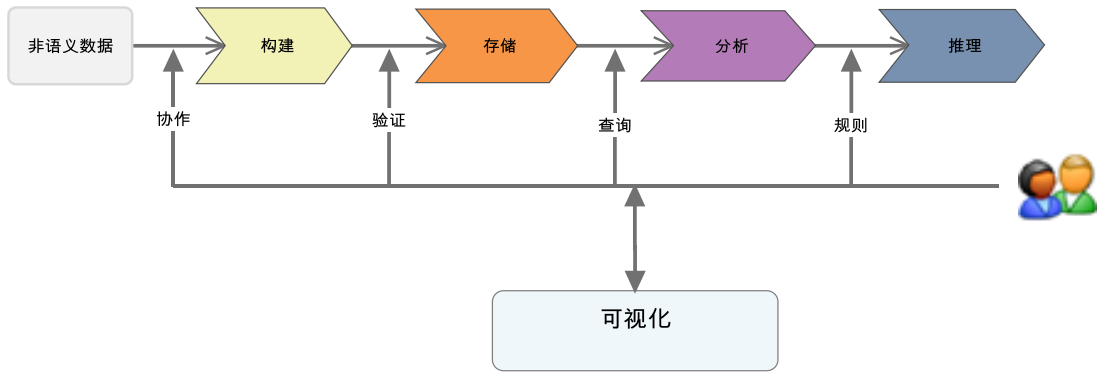
6.1.6. 各种方法比较

可视化技术	优势	不足
缩进列表	简单、熟悉，no label overlap	树形结构而不是图，只能表达继承关系，而无法描述角色联系
节点链接图和树	层次结构的全局视图	空间利用率较低，节点很多后容易造成视觉混乱
缩放图	浏览定位到特定节点很方便	无法有效看到全局视图，用户容易遗忘浏览路径，应该提供浏览的历史信息
空间填充	通过颜色编码，很容易分辨属性与类，容易形成总体的pattern	不利于大量节点可视化
上下文+焦点	节点位置可变，中心节点信息详细，全局视图	节点间的关系表示不明显

好的语义网可视化应支持下列可视化方法: **overview, zoom, filter, detail-on-demand, relate** (关系), **History** (浏览历史以支持 **undo, replay** 和 **refine**), **extract** (抽取子集合)。

可视化应该与有效的搜索工具和查询机制紧密结合，浏览并不能有效定位特定的类或实例。用户希望看到有序而清晰的信息表示结果，可视化应当利用信息上下文甚至用户的个人资料来支持本体探索。

7. 研究方向



从总体上看，语义网可分为四个过程：构建、存储、分析和推理。由于各种原因，在每个阶段都需要用户一定程度的干预。可视化在人机交互过程中能帮助用户理解语义知识，验证过

程的正确性。

在构建本体过程中，不同领域的专家需要相互协作完成。即使是构建完成后，仍需要验证结果的正确性才能存储。语义网分析是其存在的主要意义，查询是分析阶段的基础。由于语义网的复杂性，查询需要与可视化相结合才能便于用户有效得到想要的结果。利用领域规则进行推理是语义网应用的高级阶段，同时也是最复杂的过程，用户的适当参与能够验证推理的正确性。

因此，语义网的可视化在多用户协同构建本体库、本体查询、语义网分析和推理上都存在值得研究的问题。最近，随着社交网络的兴起，社会语义网分析正成为研究热点。

## 8. 社会语义网可视化方法

**Sociogram** 是最早对社会网络可视化的方法。基于语义网的社会网络分析能够有效分析社会网络中存在的各类问题，因此语义网可视化方法正被应用于该领域。（需要进一步探索）